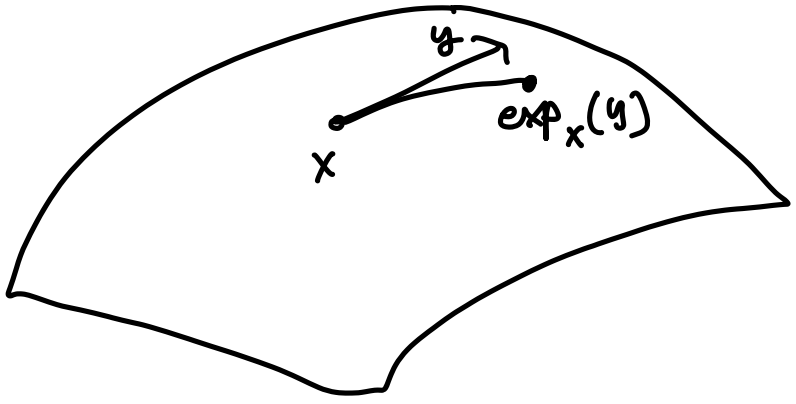
Query complexity of Riemannian optimization

Outline: 1) Quick Riemann opt. refresher
2) Examples; hyperbolic plane
3) Upper bounds
4) Hamilton-Moitra lower bound
5) Open directions

Riemannian Opt: M Riemannian manifold

- metric $\|\cdot\|_x$
- $d(x, x') :=$ length of shortest $x \to x'$ curve
  $=$ length of geodesic between.



- assume we have access to the exponential map $\exp_x : T_x M \to M$;

  goes distance $\|y\|_x$ along geodesic in direction y.

- gradient $\nabla f(x) := \nabla_{y=0} f \circ \exp_x(y)$

  "direction of steepest geodesic".

- First order optimization:

  query at x, get $f(x), \nabla f(x)$

- want to find point x s.t.

$\|\nabla_x f\| < \varepsilon$, $|f(x) - OPT| < \varepsilon$, or $|d(x, x^*)| < \varepsilon$.

- convex: $\partial_t^2 f \circ \exp_x(yt) \geq 0 \quad \forall y, \forall x \in M$
- α-strongly convex: $\partial_t^2 f \circ \exp_x(yt) \geq \alpha \|y\|_x^2$.
- β-smooth: $\partial_t^2 f \circ \exp_x(yt) \leq \beta \|y\|_x^2$

Manifold examples:

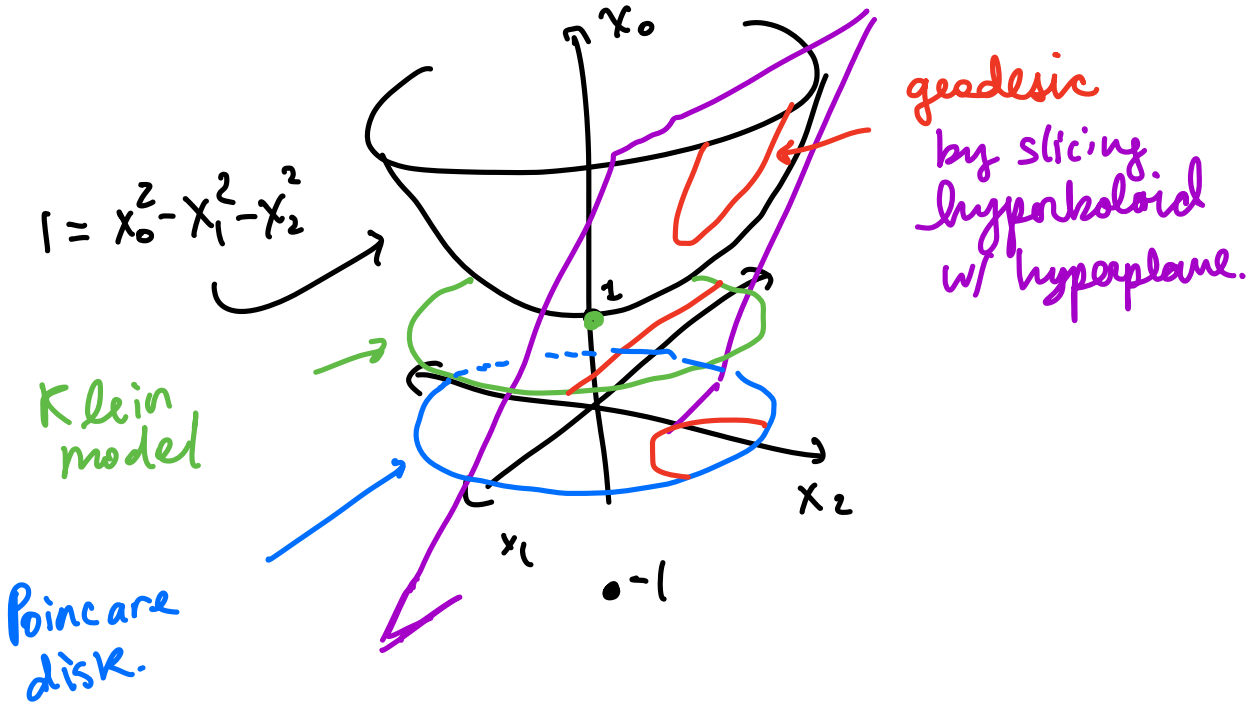
- $PD(n)$, $\exp_x(Y) = \sqrt{x} e^{Y} \sqrt{x}$. $\mathbb{R}$ or $\mathbb{C}$

- $PD_1(n) = PD(n) \cap SL(n)$
- **Hyperbolic plane**: $PD_1(n, \mathbb{R})$.

$$\det \begin{bmatrix} x_0 + x_1 & x_2 \\ x_2 & x_0 - x_1 \end{bmatrix} = 1, \iff x_0^2 - x_1^2 - x_2^2 = 1$$

$x_0 > 0$

"hyperboloid model"



Function examples

- **Kronecker covariance estimation**:

Given tensor $V \in \mathbb{R}^{d^3}$, minimize

$$\| X_1 \otimes X_2 \otimes X_3 V \| \quad \text{over } PD_1(d)^3.$$

AKA "tensor scaling".

- **Matrix means**: given $P_1, \dots P_n$ in $PD(d)$,

  minimize $\sum_i d(X, P_i)^2$ over $PD(d)$.

Examples so far are

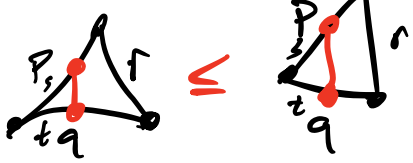
Hadamard manifolds, i.e. Riem. manifolds that:

- Have sectional curvature ≤ 0 &
- Are simply connected.

↳ $\exp_x$ of 2-D subspaces look like pringles

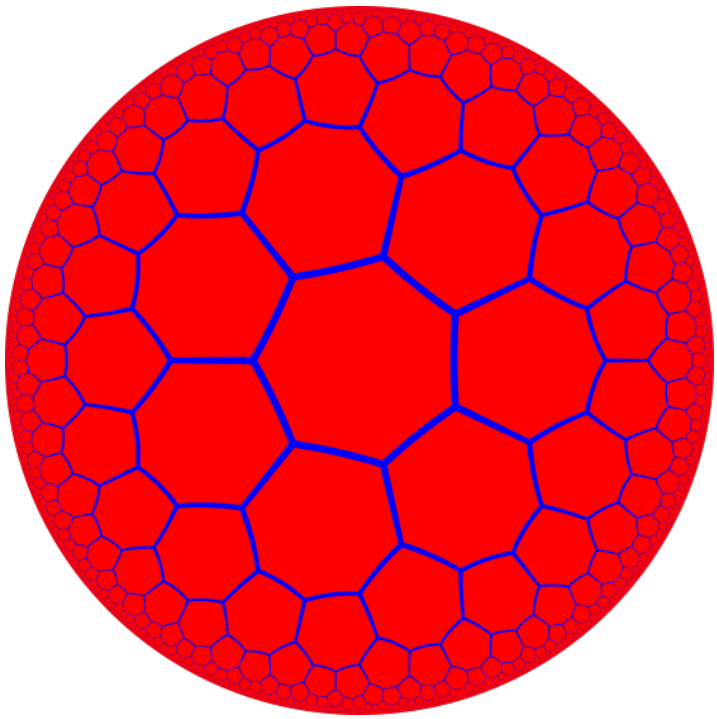
Properties: if M Hadamard, then:

- Triangle comparison:



- **Hadamard-Cartan:** $\exp_x : \mathbb{R}^n \to M$ diffeomorphism.
- **Bishop-Gromov:** $\mathrm{vol}(B_\varepsilon \subseteq M) \geq \mathrm{vol}(B_\varepsilon \subseteq \mathbb{R}^n)$
- **Distance-convexity:** $d(x,\cdot)^2$ strictly convex.

Key fact about H_2: $\mathrm{vol}\, B_x(r) = e^{\Theta(r)}$.



Poincare disk model.

Positive algorithmic results:

Parameters: β smoothness, κ condition # $\frac{\alpha}{\beta}$

R, radius of ball $\supseteq$ domain.

T = # steps, n dimension.

Class	Algorithm	Guarantee
β-smooth	GD Folklore	$\|\nabla f\| \leq \sqrt{\frac{\beta(f(x_0)-f(x^*))}{T}}$
		$f(x)-f(x^*) \leq \frac{\beta R^2}{T}$
	AGD Martinez-Rubio 2020	$f(x)-f(x^*) \leq \frac{\beta}{T^2} e^{O(R)}$ * vs $\frac{LR^2}{T^2}$ Euc.
κ-conditioned,	GD	$\|\nabla f(x)\| \leq \|\nabla f(x_0)\| e^{-T/\kappa}$
	Zhang, Sra → 2016	$f(x)-f(x^*) \leq \beta R^2 e^{-T/\kappa}$
	AGD Martinez-Rubio 2020	$f(x)-f(x^*) \leq \beta e^R e^{-T/\sqrt{\kappa}}$ * v.s. $LR^2 e^{-T/\sqrt{\kappa}}$ Euc.

* Only on H_n.

L lipschitz, differentiable	Cutting Planes Rusciano '18 Non-algorithmic!	$f(x) - f(x^*)$ $\leq nLe^{nR}e^{-T/n^2}$ volume of ball $e^{n}e^{-T}$ Euc. $LRe^{n}e^{-T}$.

Issues: Even in H_2, to get

- AGD needs $R\sqrt{K}$ queries
  vs $\sqrt{K}\log R$ Euclidean
- Cutting planes needs R
  vs $\log R$ Euclidean.

Question: Do Euclidean bounds carry over??

Answer: Probably not.

Work of Hamilton-Moitra 2020 suggests:

Conjecture:

- For 1 lipschitz, need $\Omega(R)$.
- For $O(R)$ conditioned, need $\Omega(R)$. (not $\sqrt{R}\log R$.)

$\sqrt{\kappa}\sqrt{R}$?

Comment: $O(R)$ is best possible condition # on $B(R) \subseteq H_2$. (HM13).

- They prove this in a noisy model.
- Noise can be R^{-100}; wouldn't cause problems in $\mathbb{R}^4$.
- **Why do I care?** In tensor scaling problem, need $R = \exp(n)$!

Noisy Query Model:

- Unknown differentiable function f w/ minimum x^* in $B(R)$.
- Can query any point $x \in B_{1000R}(0)$, get $f(x) + z_0$, $\nabla f(x) + z_1$ where z_0, z_1 are independent (& do not depend on past queries).
- Say z (noise) is **c-nonconcentrated** if the density function of z is everywhere $\leq c$, **δ-precise** if $\|z\| \leq \delta$ always.

- **Concretely:** let $Z_2 \sim \text{unif } B_{\mathbb{R}^2}(\delta)$; $c = \frac{1}{\pi\delta^2}$

  $Z_1 = \text{unif } B_{\mathbb{R}}(\delta)$; $c = \frac{1}{2\delta}$.

KEY FACT: Small (precise) noise doesn't interfere w/ Euclidean optimization.

Main Theorem There are function classes

- $\mathcal{F}_c$ $O(R)$-conditioned
- $\mathcal{F}_L$ $O(1)$-Lipschitz

for which finding a point $\leq \frac{R}{5}$ from x^* requires

$$\Omega\left(\frac{R}{\log R + \log\delta + \log c}\right)$$

δ-precise, c-noncone. noisy queries.

e.g. $\delta = R^{-10}$, $c = R^{20}$. $\rightsquigarrow \Omega\left(\frac{R}{\log R}\right)$.

Remarks:

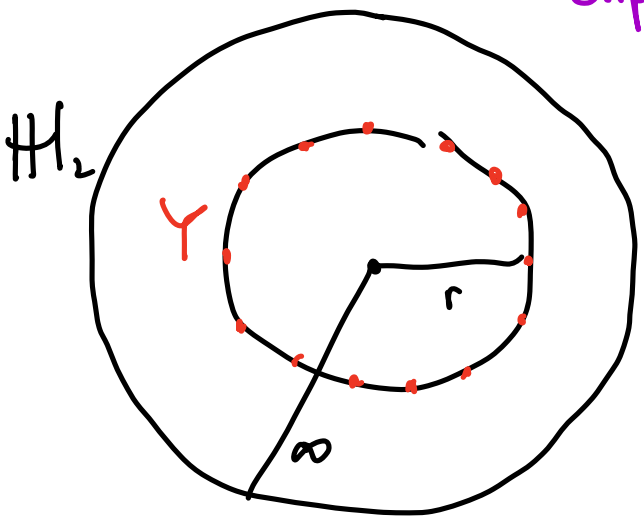
- Noise doesn't preclude $\log(\frac{1}{\varepsilon})$, just $\log(R)$, $\sqrt{R}$ dependence.
- can replace $d(x, x^*) \le \frac{R}{5}$ by $f(x) - f(x^*) \le \frac{R^2}{5}$ or $\|\nabla f(x)\| \le \frac{R}{5}$.
- many classes would work.
- also works for $\nabla^2 f$ queries.

Proof of HM'21

- **function class:** class consists of functions f_y, $f_y(x) = d^2(x, y)$. (for $k \le r$),

or $f_y(x) = d(x,y)^*$ (for lipschitz).

- **Noise necessary:** else gradient points exactly towards $y = \text{argmin}_x f_y(x)$!
- **Which y?** y is chosen from a known set Y of $n = e^{\Theta(R)}$ equally spaced points at distance $\frac{R}{2}$ around unit circle in $\mathbb{H}_2$. Can fit b/c vol exponential in R!



Figuring out which element $y \in Y \iff$
Finding point within $\frac{R}{4}$ of argmin f.

Game formulation:

- Alice knows an element $y \in Y$, $|Y| = n$.

- Bob makes a query $q \in Q := B_{1000R}(0)$

  recieves observation $x \in X$,

  Sample from distribution

$P_{q,y}$:= density of pair $\left(f(q) + z_1, \nabla f_y(q) + z_2\right)$ ← noise.

here $x \in X := [0, (1000R + \delta)^2] \times B_{\mathbb{R}^2}(1000R + \delta)$

- Bob wins if he guesses y w/ probability $\geq 2/3$.

[HM13] proves a general lower bound on noisy query games:

Theorem: For any **noisy query game**

(Y, Q, X, P) — map $p: Q \times Y \to \mathbb{P}(X)$

(Y: Alice's secrets, Q: queries, X: observations)

such that $P_{x,y}$ is c-nonconcentrated, then Bob needs $\Omega\left(\frac{\log|Y|}{\log(c\,\mathrm{vol}(X))}\right)$ queries to win.

Corollary: For noisy gradient task,

$|Y| = e^{\Theta(R)}$, $\mathrm{vol}(X) = O\left((r^2+2\delta)(r+\delta)^2\right)$.

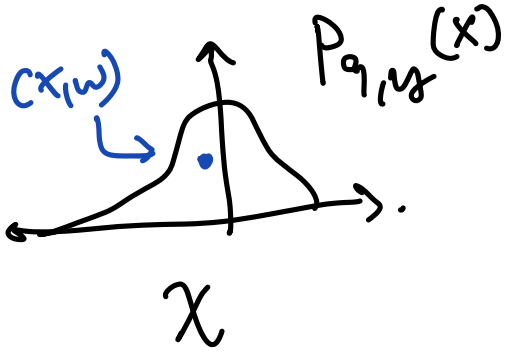
$\Rightarrow \mathrm{vol}(X) \leq O(r^3\delta^3)$;

$\Rightarrow \Omega\left(\frac{R}{\log R + \log\delta + \log c}\right)$ queries necessary.

Proof of NQG theorem:

- prove lower bound for game where Bob gets even more info.

Def: Transparent NQG: Same as NQG, except Bob gets sample from uniformly random point (x,w) under graph of $P_{q,y}: \mathcal{X} \to \mathbb{R}$



marginal distrib of x is $P_{q,y}$; only gives Bob more info.

If Alice samples $y \sim \text{Unif}(Y)$, Bob's optimal strategy after t steps is to output y^* most likely under posterior

$$y \mid \underbrace{q_1, (x_1, w_1), q_2, \dots q_t, (x_t, w_t)}_{\text{all outcomes so far}}$$

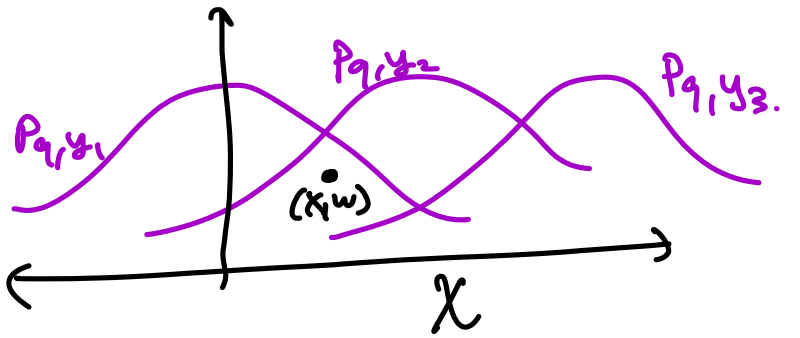
Lemma: Let $y \sim \text{Unif}(Y)$, consider TNQG.

1) posterior for y after 1 query is uniform on some $Y_1 \subseteq Y$
2) in expectation over observation,

$$\log|Y_1| \geq \log|Y| - \log(\text{Vol}(X)).$$

Proof: 1) After query q, (x, w) sampled at uniform from

graph of $P_{q,y}$ for $y \sim \text{Unif}(Y)$.



Claim: $Y_1 = \{y' : (x,w) \in \text{Graph } P_{x,y'}\}$.

- After query, Bob knows $y \neq y'$ for any $y' \notin Y_1$.
- Moreover, because every graph has area 1, posterior is uniform on all $y' \in Y_1$.

2) Let $N(x,w) := |Y_1|$. Then

$$\mathbb{E}_y \left[\int_{\text{Graph } P_{q,y}} \log N(x,w)\, dx\, dw \right]$$

$$= \mathbb{E}_{\text{noncom. } \gamma}\left[\int_{\chi \times [0,c]} 1_{\text{graph } P_{q,w}}(x,w) \log N(x,w)\, dx\, dw \right]$$

$$= \int_{\chi \times [0,c]} \frac{N(x,w)}{|\gamma_1|} \log N(x,w)\, dx\, dw \quad *$$

- Each graph has area $1 \Rightarrow \int_{\chi \times [0,c]} \frac{N(x,w)}{|\gamma|} dx dw = 1.$

- Thus $*$ is $-$ entropy of $N(x,w)/|\gamma_1|$ up to additive const; $\Rightarrow *$ minimized when $N(x,w) \equiv$ const. over $\chi \times [0,c]$. $\equiv \frac{|\gamma_1|}{c\, \mathrm{vol}(\chi)}$.

- Min value for $*$:

$$\int \frac{1}{c\,\mathrm{vol}(\chi)} \log\left(\frac{|\gamma|}{c\,\mathrm{vol}(\chi)}\right) dx dw$$

$$= \log\left(\frac{|\gamma|}{c\,\mathrm{vol}(\chi)}\right) = \log|\gamma| - \log(c\,\mathrm{vol}(\chi)).$$

□

- **Now using Lemma, prove NQG theorem.**

- By induction, after t queries, posterior uniform on γ_t with

$$\mathbb{E} \log|\gamma_t| \geq \log|\gamma| - T\log[c\,\mathrm{vol}(X)]$$

or $Z := \log|\gamma| - \log|\gamma_t|$

has $\mathbb{E}Z \leq T\log[c\,\mathrm{vol}(X)]$.

Algorithm succeeds if $|\gamma_t| = 1$,

i.e. $Z \geq \log|\gamma|$.

By Markov, $\Pr[Z] \geq 3T\log[c\,\mathrm{vol}(X)]$

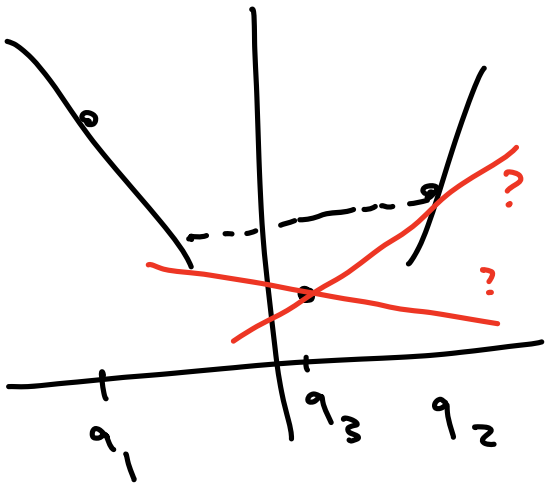
$\leq \frac{1}{3}$; hence win with

$pr \leq \frac{1}{3}$ if $T \leq \dfrac{\log|\gamma|}{3\log[c\,\mathrm{vol}(X)]}$. □.

Open directions:

- Removing noise?

  Play "20 questions": $\frac{1}{2}$ search space



Issue: which gradients consistent w/ previous observations?

Nontrivial in H_2; no "linear functions"

$V \cdot \log_x(y)$ not convex in y ∴

- Algorithmic cutting planes / ellipsoid in H_n ... H_2 maybe O.K.

- Function classes that can be optimized in $O(\log R)$ queries? & include interesting problems like tensor scaling ☺